The distribution of calibrated likelihood-ratios in speaker recognition

David A. van Leeuwen[1] *and Niko Brümmer*[2]

[1]Netherlands Forensic Institute, The Hague and Radboud University Nijmegen, The Netherlands
[2]AGNITIO Research, Somerset West, South Africa

Abstract

This paper studies properties of the score distributions of calibrated log-likelihood-ratios that are used in automatic speaker recognition. We derive the essential condition for calibration that the log likelihood ratio of the log-likelihood-ratio is the log-likelihood-ratio. We then investigate what the consequence of this condition is to the probability density functions (PDFs) of the log-likelihood-ratio score. We show that if the PDF of the non-target distribution is Gaussian, then the PDF of the target distribution must be Gaussian as well. The means and variances of these two PDFs are interrelated, and determined completely by the discrimination performance of the recognizer characterized by the equal error rate. These relations allow for a new way of computing the offset and scaling parameters for linear calibration, and we derive closed-form expressions for these and show that for modern i-vector systems with PLDA scoring this leads to good calibration, comparable to traditional logistic regression, over a wide range of system performance.

1. Introduction

In recent years, calibration in automatic speaker recognition has received more attention [1–11]. Intuitively, calibration is related to the ability to properly set a threshold in a speaker detection system so as to minimize the expected error [12]. In speaker detection, the task is to decide whether or not two speech signals originate from the same speaker. Because all speaker recognition systems internally work with some scalar *score* that expresses speaker similarity, a score threshold can control the trade-off between the two types of errors that a system can make [13, 14]. Indeed, in the series of NIST Speaker Recognition Evaluations (SRE) the primary evaluation measure has been sensitive to calibration. Until SRE 2010, calibration was assessed in a single operating point, through a single decision cost function known as C_{det}. Also other technologies in speech technology or biometrics utilize calibration-sensitive evaluation measures, such as the cost functions C_{avg} in language recognition [15] and the Half Total Error Rate in face recognition [16].

Since around 2004 [1,2] the concept of calibration in speaker recognition has been generalized to a range of operating points by using proper scoring rules [17] to evaluate probabilistic statements about whether a trial is a same-speaker (target) or different-speaker (non-target) trial. A system that represents its score as a *likelihood-ratio* can be well-calibrated over a wide range of operating points simultaneously. This representation of the speaker recognition score has direct application in speaker detection, as the decision threshold follows directly from the cost function parameters [14], but also in evidence reporting in forensic speaker comparison cases [4, 18]. In the NIST SRE 2012, for the first time, hard decisions were no longer required, and instead the recognition score had to be submitted in the form of a likelihood-ratio. The evaluation measure effectively sampled the decision cost function at two different parameters [19, 20].

Since a calibrated likelihood-ratio is still just a score, all properties of normal scores apply to likelihood-ratios as well, and we can draw DET and ROC plots, determine EERs and inspect the score distributions. The axis warping of the DET plot [13] in combination with the observed more-or-less straight DET curves suggests that target and non-target score distributions could be accurately modelled with Gaussians. These score distributions and the relation to the DET have been studied previously [21, 22] and are very instructive to the understanding of basic detection theory and the concepts of calibration [14, 23]. In this paper we are interested in properties of the distributions of *calibrated log-likelihood-ratios*. This may help situations were we carry out a calibration transformation on raw recognition scores, because it can tell us what the calibrated distributions should look like.

The paper is organized as follows. We define the very nature of a calibrated likelihood-ratio in Section 2. In Section 3 we investigate the properties of log-likelihood-ratio distributions when they are Gaussian, and we will then apply these in Section 4 as a new method for calibration. We then present experiments and conclusions.

2. Likelihood-ratio idempotence

Here we carefully define the *likelihood-ratio* (LR) and show that it has the interesting property: *the LR of the LR is the LR*, which forms a definition of calibration.

The speaker recognition system has as input two speech segments, denoted X and Y, which it processes in two steps. We represent the first step as $s = f(X, Y)$. To keep things general, s may represent different kinds of output, e.g., a pair of acoustic feature vector sequences, a pair of i-vectors, or just a single, scalar recognition score. The second step is to compute the likelihood-ratio r as a function of s, as:

$$r = \frac{P(s \mid H_1, \mathcal{M})}{P(s \mid H_2, \mathcal{M})} \tag{1}$$

where H_1 is the (target) hypothesis that X and Y originate from the same speaker, H_2 the (non-target) hypothesis that they are from two different speakers, and $\mathcal{M}$ is a generative probabilistic model for s. In current practice, s is always the recognition score, so that $\mathcal{M}$ merely models scalar scores—not i-vectors, acoustic feature sequences or speech signals. But our theory

below is sufficiently general to remain applicable in future to more ambitious models, when s might have a more complex form. We now assume there is given the *hypothesis prior*, $\pi = P(H_1)$, which allows us to express the *hypothesis posterior*, via Bayes' rule as:

$$P(H_1 \mid s, \mathcal{M}, \pi) = \frac{\pi r}{\pi r + (1-\pi)} \tag{2}$$

This shows that r is a *sufficient statistic*: the posterior depends on s only through r. This allows rewriting the posterior as:

$$P(h \mid s, \mathcal{M}, \pi) = P(h \mid r, \mathcal{M}', \pi), \qquad h \in \{H_1, H_2\} \tag{3}$$

where we have introduced $\mathcal{M}'$ to denote $\mathcal{M}$, augmented by asserting (1). Although r contains all the relevant information that $\mathcal{M}$ can extract from s to recognize the unknown hypothesis, it must be stressed that r and s do *not* necessarily contain all the relevant information that could have been extracted from the original input X, Y by some more elaborate model. Now we use the *odds form* of Bayes' rule:

$$\frac{P(H_1 \mid \rho, M, \pi)}{P(H_2 \mid \rho, M, \pi)} = \frac{\pi}{1-\pi} \frac{P(\rho \mid H_1, M)}{P(\rho \mid H_2, M)} \tag{4}$$

where ρ is a placeholder for r or s and M for $\mathcal{M}$ or $\mathcal{M}'$. Combining this with (3), we find the desired relationship (the LR of the LR is the LR [24]):

$$r = \frac{P(s \mid H_1, \mathcal{M})}{P(s \mid H_2, \mathcal{M})} = \frac{P(r \mid H_1, \mathcal{M}')}{P(r \mid H_2, \mathcal{M}')}. \tag{5}$$

If we define x to be the *log-likelihood-ratio* (LLR):

$$x = \log r \tag{6}$$

we also find[1] (the LLR of the LLR is the LLR):

$$x = \log \frac{P(x \mid H_1, \mathcal{M}'')}{P(x \mid H_2, \mathcal{M}'')} \tag{7}$$

where $\mathcal{M}''$ augments $\mathcal{M}'$ by addition of (6).

2.1. Implications

Rewriting (5) as:

$$P(r \mid H_1, \mathcal{M}') = rP(r \mid H_2, \mathcal{M}') \tag{8}$$

we see that if either of the two distributions is given, then the other distribution is completely determined—they cannot vary independently. Moreover, a further restriction is placed on these distributions: since the LHS must integrate to 1, the *expected value* of the non-target distribution (the integral of the RHS) must be: $\langle r \rangle = 1$. Similarly, for targets: $\langle \frac{1}{r} \rangle = 1$. By applying Jensen's inequality [25] we also find for targets: $\langle x \rangle \geq 0$ and for non-targets: $\langle x \rangle \leq 0$.

2.2. Good and bad calibration

How does (5) function as a definition of calibration? Since it is an equality, won't all LRs calculated via (1) by some model $\mathcal{M}$, just automatically satisfy (5)? Yes they will, but only if $\mathcal{M}$ and $\mathcal{M}'$ are related as explained above. If we want to independently judge the goodness of the calibration of r, we do not condition the distributions for r on the recognizer's model $\mathcal{M}$. Instead, we could empirically observe the target and non-target values of r as calculated by the recognizer over an independent, supervised database of speaker detection trials. Letting $\mathcal{O}$ denote the empirical observation, we could then say the model $\mathcal{M}$ is well calibrated if:

$$r = \frac{P(s \mid H_1, \mathcal{M})}{P(s \mid H_2, \mathcal{M})} \approx \frac{P(r \mid H_1, \mathcal{O})}{P(r \mid H_2, \mathcal{O})} \tag{9}$$

Bad calibration is when the LRs given respectively by the recognizer's $\mathcal{M}$ and empirical observation $\mathcal{O}$, do not agree in this way. This can and does happen, since $\mathcal{O}$ is independent of any development data that was used to determine the form and parameters of $\mathcal{M}$.

It should be noted that (9) does not give a practical recipe to judge degree of goodness of calibration—it specifies neither how to assign $P(r \mid h, \mathcal{O})$, nor how to numerically evaluate the agreement between LHS and RHS. For practical solutions for calibration-sensitive objective functions, see for example [26].

3. Gaussian distributed log-likelihood-ratios

Inspired by the fact that DET curves in speaker recognition tend to be straight [21], we explore a Gaussian solution to the LLR distribution constraint (7). Since target and non-target LLR distributions are so tightly coupled, it turns out that if the one is assumed to be Gaussian, then the other must also be. We shall use the shorthand: $e(x) = P(x \mid H_1, \mathcal{M}'')$ and $d(x) = P(x \mid H_2, \mathcal{M}'')$. Arbitrarily assuming a Gaussian distribution for non-targets (*different*-speaker trials):

$$d(x) = \mathcal{N}(x \mid \mu_d, \sigma_d) = \frac{1}{\sqrt{2\pi}\sigma_d} e^{-(x-\mu_d)^2/2\sigma_d^2}. \tag{10}$$

We derive the functional form for targets[2], $e(x)$, when (7) applies:

$$e(x) = e^x d(x) = \frac{1}{\sqrt{2\pi}\sigma_d} e^{x-(x-\mu_d)^2/2\sigma_d^2}. \tag{11}$$

We collect the terms in x in the exponent, which itself can be written like

$$-\frac{x^2 - 2\mu_d x + \mu_d^2}{2\sigma_d^2} + \frac{2\sigma_d^2 x}{2\sigma_d^2} \tag{12}$$

$$= -\frac{x^2 - 2(\mu_d + \sigma_d^2)x + \mu_d^2}{2\sigma_d^2} \tag{13}$$

$$= -\frac{\left(x - (\mu_d + \sigma_d^2)\right)^2}{2\sigma_d^2} + \frac{2\mu_d\sigma_d^2 + \sigma_d^4}{2\sigma_d^2} \tag{14}$$

The first term is in the familiar form of a Gaussian exponent, the second will result in a constant factor. Gathering terms, and writing

$$\mu_e = \mu_d + \sigma_d^2, \tag{15}$$

the expression for the same-speaker comparison log-likelihood-ratio scores becomes

$$e(x) = \frac{1}{\sqrt{2\pi}\sigma_d} e^{\sigma_d^2/2 + \mu_d} e^{-(x-\mu_e^2)/2\sigma_d^2} \tag{16}$$

$$= e^{\sigma_d^2/2 + \mu_d} \mathcal{N}(x \mid \mu_e, \sigma_d). \tag{17}$$

We see that $e(x)$ is of Gaussian shape, with

$$\sigma_e = \sigma_d \equiv \sigma. \tag{18}$$

[1] To see this, note the log transformation is monotonic and the Jacobian of the transformation cancels in the ratio.

[2] trials where the speakers are *equal*

Since $e(x)$ must be a proper PDF, its integral over x must be unity, from which follows that

$$e^{\sigma^2/2+\mu_d} \int_{-\infty}^{\infty} \mathcal{N}(x \mid \mu_e, \sigma)\, dx = 1 \tag{19}$$

$$-2\mu_d = \sigma^2. \tag{20}$$

Finally, with (15) we find

$$\mu_e = \mu_d + \sigma^2 = -\mu_d \equiv \mu, \tag{21}$$

This shows that $d(x)$ and $e(x)$ are equal variance Gaussians with means symmetric around zero at $\pm\mu$, and where the variance and mean are related (20)

$$\sigma^2 = 2\mu. \tag{22}$$

3.1. Equal Error Rate and d'

Using the symmetry of the solution, it is clear that the threshold for the equal error rate is at $x = 0$. Using the expression for the miss probability, the equal error rate $E_=$ is

$$E_= = \int_{-\infty}^{0} \mathcal{N}(x \mid \mu, \sigma) \tag{23}$$

$$= \int_{-\infty}^{-\mu/\sigma} \mathcal{N}(x \mid 0, 1) \equiv \Phi(-\mu/\sigma), \tag{24}$$

where $\Phi(x)$ is the cumulative normal distribution.

It is sometimes useful to recognize the parameter d' from detection theory, which is the difference in means expressed in terms of the standard deviation, here $d' = 2\mu/\sigma$. With (24) the relation becomes

$$E_= = \Phi(-\frac{1}{2}d'). \tag{25}$$

$$d' = \sigma = -2\Phi^{-1}(E_=), \tag{26}$$

introducing $\Phi^{-1}(y)$, the inverse of the cumulative normal distribution. The importance of the relations above is that μ and σ are determined by the discrimination performance measured by $E_=$, using (22) and (26)

$$\mu = \frac{\sigma^2}{2} = 2[\Phi^{-1}(E_=)]^2. \tag{27}$$

4. A new calibration method

In practice, automatic speaker recognition systems do not deliver scores that can directly be interpreted as a log-likelihood-ratio, even though they are computed as such, for instance in the good old UBM-GMM scoring [27] or the latest i-vector PLDA scoring [28]. A practical solution to this is to convert raw scores $s(X, Y)$ to calibrated log-likelihood-ratios by some transformation function $x(s)$, usually constrained to be monotonic increasing. There are many ways of doing this. The FoCal [29] and BOSARIS [30] toolkits use logistic regression to discriminatively train linear calibration transformations. Other possibilities include isotonic regression (PAV [30]) and line-up calibration [9] that uses the rank in a line-up of foil speakers. In FoCal or BOSARIS, the score-to-LLR function is affine:

$$x(s) = as + b \tag{28}$$

and the parameters a and b are found by optimizing cross-entropy, a calibration-sensitive objective function defined on a supervised set of speaker recognition trials.

Here we contrast the popular discriminative logistic regression solution to a new generative, constrained maximum-likelihood (ML) solution. Our constraints follow from assuming (i) Gaussian LLR distributions, and (ii) an affine score-to-LLR transform (28). This implies that (i) the LLR distributions are constrained as derived in Section 3, and (ii) the score distributions are also Gaussians, with equal variances. With no LLR distribution constraints, we would have had 6 free parameters: 2 means, 2 variances and 2 calibration parameters. But we have imposed 3 constraints, equal variances (18), symmetric means (21) and (22). We find the remaining 3 free parameters by maximizing the following weighted likelihood:

$$\frac{\alpha}{N_e} \sum_{i\in\mathcal{E}} \log \mathcal{N}(s_i \mid m_e, v) + \frac{1-\alpha}{N_d} \sum_{i\in\mathcal{D}} \log \mathcal{N}(s_i \mid m_d, v)$$

where $\mathcal{E}$ and $\mathcal{D}$ index N_e target, and N_d non-target scores, weighted by α and $1-\alpha$, respectively. The score distribution parameters that need to be optimized are the means m_e, m_d and common variance v. Setting derivatives to 0, we find the maximum likelihood at the sample means:

$$m_e = \frac{1}{N_e} \sum_{i\in\mathcal{E}} s_i, \qquad m_d = \frac{1}{N_d} \sum_{i\in\mathcal{D}} s_i \tag{29}$$

and at a weighted combination of sample variances:

$$v = \frac{\alpha}{N_e} \sum_{i\in\mathcal{E}} (s_i - m_e)^2 + \frac{1-\alpha}{N_d} \sum_{i\in\mathcal{D}} (s_i - m_d)^2 \tag{30}$$

By (28), the LLR distribution parameters become $\sigma^2 = a^2 v$, $\mu_e = am_e + b$ and $\mu_d = am_d + b$. Finally, applying the constraints $\sigma^2 = \mu_e - \mu_d$ and $\mu_e = -\mu_d$, we can solve for the calibration parameters:

$$a = \frac{m_e - m_d}{v}, \qquad b = -a\frac{m_e + m_d}{2} \tag{31}$$

We call this recipe *constrained, maximum-likelihood, Gaussian* (CMLG) calibration. An advantage of CMLG is that it has a closed form, in contrast to the iterative optimization required by logistic regression.

4.1. Experiment

In order to test CMLG we apply it to a number of recognition trials sets. We use a set of trials crafted for duration-dependence experiments [8] from the NIST SRE 2008 and 2010 trial sets, the telephone-telephone "extended" trial lists. We constructed short duration segments of 5, 10, 20, and 40 seconds from both train and test segments by simply selecting the first frames after speech activity detection. All durations, including the full conversation side, were tested in all combinations, leading to 25 different trial lists. The NIST SRE10 'det-5' performance over these lists ranges from $E_= = 2.9$–$26\,\%$. The recognition system is a standard i-vector based system with PLDA scoring described elsewhere [20].

We contrast CMLG (with $\alpha = \frac{1}{2}$) to the traditional logistic regression method. The calibrations are trained on NIST SRE 2008 data (427 375 trials) and applied to SRE 2010 trials for evaluation (10 007 900 trials), all gender mixed. We evaluate the 25 different trial list combinations using C_{llr}, a cost function that is sensitive to calibration over the whole DET curve [2]. We used R's `glm` routine for logistic regression.

The results are shown in Fig. 1, where we have plotted the C_{llr} obtained using CMLG calibration versus C_{llr} obtained using logistic regression. The values are highly correlated. For

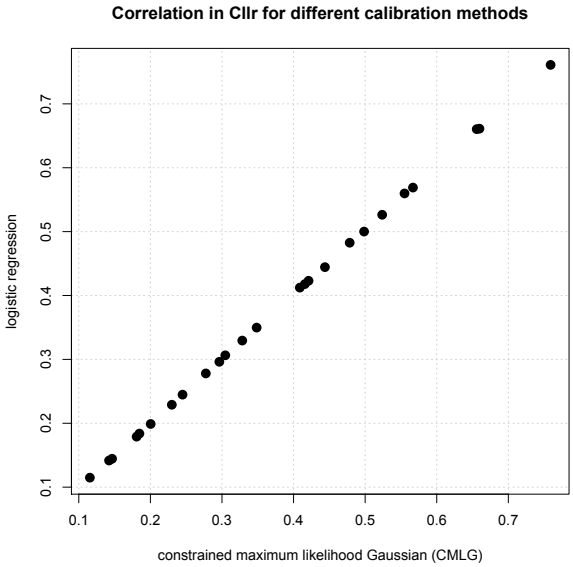

Figure 1: C_{llr} values of the 25 trial lists for the CMLG method (horizontal) versus logistic regression (vertical).

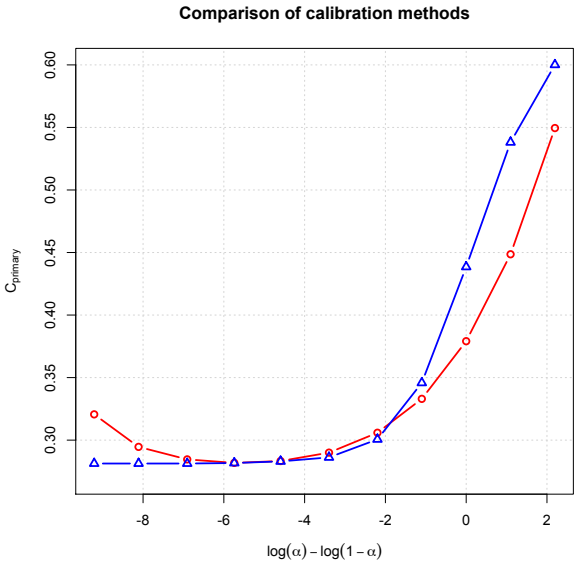

Figure 2: C_{primary} for logistic regression and CMLG calibration methods for ABC's SRE12 submission, as a function of prior α used in the objective / ML optimization.

CMLG, the average C_{llr} over all 25 conditions is 0.375, for logistic regression it is 0.376. These can be called good, as the mean $C_{\text{llr}}^{\text{min}}$ is 0.370.

We have also used the NIST SRE12 scores from the ABC-team to study the effect of α in (30) to another calibration sensitive measure C_{primary}, cf. Fig. 2, for details we refer to [26]. The figure shows that with CMLG good calibration results can be obtained for a different system with different data and a different performance measure, if the correct α is chosen.

5. Discussion and Conclusions

We have shown in this paper, that if the different-speaker calibrated log-likelihood-ratio scores from a speaker recognition system follow a Gaussian distribution, then the distribution of the same-speaker scores must also be Gaussian after calibration, with the same variance but opposite mean. Because monotonically increasing score-to-likelihood-ratio functions do not change the DET plot, such equal-variance distributions in the calibrated score domain imply 45° DET-plots in the raw score domain as well—which is neither observed with real data[3] nor desired for applications operating in the low false alarm region. The logical conclusion then is that real scores, if they are well-calibrated, will not be Gaussian. However, we see that our PLDA system can be calibrated quite well under the Gaussian assumptions, and indeed we have noticed that i-vector PLDA systems tend to have score distributions that appear more Gaussian than earlier technologies, such as i-vector LDA cosine distance scoring, support vector machines or the UBM-GMM likelihood ratio scoring.

The Gaussian solution to the LLR equation (7) is one where both distributions are shaped by the same mathematical function. In signal detection theory, where the distribution represents noise, this seems almost mandatory, but in speaker recognition this is not an obvious assumption. We have experimented with other distributions, e.g., in the likelihood-ratio domain (5) a pair of Gamma distributions is a solution to the calibration condition, and these are asymmetric in the log-likelihood-ratio domain. However, such distributions seem to be not at all representative of real score distributions. Also, an arbitrary linear combination of Gaussians with different means and corresponding variances is a solution to (7) which allows some freedom in fitting a shape of score distribution. In principle, there is no need for real score distributions to follow any mathematical description, but we have observed that many researchers like to use some form of idealized shape of the score distributions to understand the data [4, 21]. When calibration methods are designed, condition (7) should therefore be taken into account.

The relations derived in Section 3 open up more possibilities for relations between the various evaluation measures. For instance, we can compute C_{llr} by numerical integration as

$$\frac{1}{\log 2}\int_{-\infty}^{\infty}\mathcal{N}(x\mid\mu,\sigma)\log(1+e^{-x})\,dx \qquad (32)$$

and this relates C_{llr} to $E_=$ via (26) and (27) for Gaussian score distributions. E.g., for our set of 25 trial lists this expression differs from $C_{\text{llr}}^{\text{min}}$ only 0.006 in root mean squared difference, or about 2 %. Instead of for calibration, the relations can also be used for fusion of systems. For pre-calibrated systems this leads to solutions that transparently depend on the correlation between the scores.

The fact that we can obtain the linear calibration parameters under the Gaussian assumption is an interesting side-effect of this study. The calibration parameters can be expressed in closed-form, and do not explicitly consider cross entropy or C_{llr} as an optimization objective. For score distributions that do not resemble a Gaussian, this calibration method is likely to fail—we therefore do not recommend CMLG calibration as a general technique. Still, we are quite pleased that the experiments support the mostly theoretical results of this paper.

6. Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Frame work Programme (FP7/2007–2013) under grant agreement no. 238803.

[3] We have measured the slope of the DET in the conventional error region 0.1–50 % for the data in the experiment. The mean slope over the 25 conditions is -0.99 with a standard deviation of 0.06, so in fact this data appears to honour the equal variance condition quite well.

7. References

[1] N. Brümmer, "Application-independent evaluation of speaker detection," in *Proc. Odyssey 2004 Speaker and Language recognition workshop*. ISCA, June 2004, pp. 33–40.

[2] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, pp. 230–275, 2006.

[3] N. Brümmer and D. A. van Leeuwen, "On calibration of language recognition scores," in *Proc. Odyssey 2006 Speaker and Language recognition workshop*, San Juan, June 2006.

[4] D. Ramos-Castro, J. González-Rodríguez, and J. Ortega-Garcia, "Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework," in *Proc. Odyssey 2006 Speaker and Language Recognition Workshop*, 2006.

[5] D. Ramos, "Forensic evaluation of the evidence using automatic speaker recognition systems," Ph.D. dissertation, Universidad Autonoma de Madrid, November 2007.

[6] N. Brümmer, L. Burget, J. Černocký, O. Glembek, F. Grezl, M. Karafiát, P. Matějka, D. A. van Leeuwen, P. Schwarz, and A. Strassheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Speech, Audio and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[7] Z. Jancik, O. Plchot, N. Brummer, L. Burget, O. Glembek, V. Hubeika, M. Karafiát, P. Matejka, T. Mikolov, A. Strasheim *et al.*, "Data selection and calibration issues in automatic language recognition–investigation with BUT-AGNITIO NIST LRE 2009 system," in *Proc. Speaker and Language Odyssey*, 2010.

[8] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "Evaluation of i-vector speaker recognition systems for forensic application," in *Proc. Insterspeech*. Firenze: ISCA, August 2011.

[9] D. A. van Leeuwen and N. Brümmer, "A speaker line-up for the likelihood ratio," in *Proc. Interspeech*. Firenze: ISCA, August 2011.

[10] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *Proc. ICASSP*. Kyoto: IEEE, March 2012.

[11] G. R. Doddington, "The role of score calibration in speaker recognition," in *Proc. Interspeech*, 2012.

[12] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation—Overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, pp. 225–254, 2000.

[13] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech 1997*, Rhodes, Greece, 1997, pp. 1895–1898.

[14] D. A. van Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," in *Speaker Classification*, ser. Lecture Notes in Computer Science / Artificial Intelligence, C. Müller, Ed. Springer, 2007, vol. 4343.

[15] A. F. Martin and A. N. Le, "NIST 2007 language recognition evaluation," in *Proc. Speaker and Language Odyssey*. Stellenbosch, South Afrika: IEEE, 2008.

[16] R. Wallace, M. McLaren, C. McCool, and S. Marcel, "Cross-pollination of normalization techniques from speaker to face authentication using gaussian mixture models," *Information Forensics and Security, IEEE Transactions on*, vol. 7, no. 2, pp. 553–562, 2012.

[17] M. DeGroot and S. Fienberg, "The comparison and evaluation of forecasters," *The Statistician*, pp. 12–22, 1983.

[18] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2104–2115, September 2007.

[19] C. S. Greenberg, "The NIST year 2012 speaker recognition evaluation plan," 2012. [Online]. Available: http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf

[20] D. A. van Leeuwen and R. Saeidi, "Knowing the non-target speakers: the effect of the i-vector population for PLDA training in speaker recognition," in *Proc ICASSP*. Vancouver: IEEE, 2013.

[21] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[22] J. Navrátil and G. N. Ramsawamy, "The awe and mistery of t-norm," in *Proc. Eurospeech*, 2003, pp. 2009–2012.

[23] N. Brümmer, "Measuring, refining and calibrating speaker and language information extracted from speech," Ph.D. dissertation, Stellenbosch University, 2010.

[24] K. Slooten and R. Meester, "Forensic identification: Database likelihood ratios and familial DNA searching," *arXiv:1201.4261 [stat.AP]*, 2012.

[25] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, vol. 30, no. 1, pp. 175–193, 1906.

[26] N. Brümmer and G. Doddington, "Likelihood-ratio calibration using prior-weighted proper scoring rules," in *Proc. Interspeech*. ISCA, 2013.

[27] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[28] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.

[29] N. Brümmer, *FoCal-II: Toolkit for calibration of multi-class recognition scores*, August 2006, software available at http://www.dsp.sun.ac.za/~nbrummer/focal/index.htm.

[30] E. de Villiers and N. Brümmer, *The Bosaris Toolkit*, BOSARIS, 2010, software available at https://sites.google.com/site/bosaristoolkit/.